4/29/25 8:45am

# Animal Pose Estimation: Cross-Species Data and Keypoint Schemes


Shaan Chanchani
Department of Engineering
Purdue University
West Lafayette, USA
schancha@purdue.edu

Claire Kim
Department of Engineering
Purdue University
West Lafayette, USA
kim3386@purdue.edu

Josh Mansky
Department of Engineering
Purdue University
West Lafayette, USA
jmansky@purdue.edu

Zian Pan
Department of Engineering
Purdue University
West Lafayette, USA
pan385@purdue.edu

Medhashree Parhy
Department of Computer Science
Purdue University
West Lafayette, USA
mparhy@purdue.edu

Armaan Sayyad
Department of Data Science
Purdue University
West Lafayette, USA
asayyad@purdue.edu

Parth Thakre
Department of Computer Science
Purdue University
West Lafayette, USA
pthakre@purdue.edu


reference Lindshield

different word choice due to community usage


***Abstract*—Our research aims to increase the robustness and efficiency of antelope pose estimation models through refining the ground-truth labels of keypoints and compressing the training dataset by filtering for similar species.**

**We focus on AP-10K [1], a prominent animal pose dataset, which we observe to have inconsistent keypoint definitions across images, hindering model performance. To address this, we create two distinct keypoint definitions. First, in an effort to have more consistently labeled images, we create a "Visible" keypoint definition, in which we choose the most visibly apparent features, enabling easier and more consistent labeling. Second, we focus on the biologically accurate point for keypoints, which would be harder for labelers, but could allow the model to better generalize on the features.**

**We utilize three methods to compress an off-the-shelf pose estimation model's training data to only include species morphologically similar to antelopes. We evaluate morphological similarity of animals with three metrics: the normalized variance of each keypoint's distance from the animals' centroid, the animals' normalized limb lengths, and the visual feature embeddings extracted by Meta's DINOv2 transformer model.**

**Through our efforts to improve model performance in keypoint estimation for specifically antelopes, we contribute to improving keypoint estimation for all animals.**


baseline of ...

## I. Introduction

Conserving biodiversity hinges on our ability to monitor wildlife populations and understand interspecies relationships. In West Africa, one such relationship of interest is between chimpanzees and their prey species, particularly antelopes. Monitoring antelope behavior and abundance offers valuable insights into chimpanzee survival strategies and broader ecosystem health.

However, obtaining high-quality, labeled data for such non-domesticated species is a persistent challenge. Pose estimation offers a scalable solution, enabling automated detection and skeletal mapping of animals across camera trap imagery. This capability supports downstream applications like behavior classification, health monitoring, and predator-prey interaction analysis.

In this work, we enhance pose estimation performance for antelopes by refining two key components. First, we propose a biologically informed keypoint labeling scheme to improve annotation consistency across low-quality images. Second, we evaluate model performance when trained on subsets of morphologically similar species selected from AP-10K [1] using novel quantitative similarity metrics. that

Our results demonstrate ~~how~~ carefully defined keypoints and biologically grounded dataset curation can improve generalization to unseen, low-resource wildlife imagery, which is an essential step toward deploying pose estimation in conservation research. Need similar conclusion for second half.


Authors ordered in alphabetical order by last name. With thanks to Haoyu Chen and Dr. Amy Reibman.


## II. Related Work

*AP-10K*

Pose estimation has primarily focused on human subjects, however, the challenge of accurately estimating animal poses has received less attention. AP-10K [1] addresses this gap by

Also useful to describe sections + where reader will find what info

offering a large-scale dataset for animal pose estimation with over 10,000 images spanning 54 species. The paper uses three methods to evaluate the dataset: supervised learning, cross-domain transfer from human pose estimation, and domain generalization. AP-10K [1] has proven to be a valuable resource for improving the accuracy and robustness of animal pose estimation models, providing new avenues for research that can extend to diverse applications such as wildlife tracking and behavioral analysis.

*Animal Pose*

In addition to AP-10K, Cao et al [2] proposed a Cross-Domain Adaptation framework specifically geared towards animal pose estimation. This work points out that when directly migrating a human pose estimation model to animals, the model performance is usually degraded due to domain differences in the distribution of appearance features, body structures, and keypoints. To address this problem, the authors propose a combined weakly supervised and semi-supervised cross-domain adaptation (WS-CDA) strategy, which utilizes a small amount of labeled animal data and a large number of unlabeled animal images to continually optimize the model by generating pseudo-labels and employing incremental training. By bridging the domain gap between humans and animals, the method greatly improves the performance of animal pose estimation compared to direct fine-tuning. This work emphasizes the importance of domain adaptivity and inspires subsequent research, including our work in this project, to place more emphasis on training data selection and improvement of keypoint labeling specifications when performing pose estimation for specific animal species.

*OpenApePose*

Additionally, OpenApePose [3] highlighted that large, tailored, species-specific annotated sets are currently still superior to larger multi-species sets, as current models have a limited capacity of generalizing across species, even within the same taxonomic species. They further emphasized this point by showing that the model trained on all ape species except the one being tested outperformed the full(more training data) OpenMonkeyPose model[LINK to paper], highlighting that even the close phylogenetic relationship between monkeys and apes does not seem to aid in keypoint estimation performance. These two key results informed our exploration into generating training datasets for a species-specific estimator based of off taxonomic relations and using different quantitative techniques to assess visual relations, to develop a quantitative method for determining the optimal data for training species-specific keypoint estimators.

### A. 300 Faces

The 300 Faces In-The-Wild Challenge [4] introduced a unified benchmark for facial landmark localization under unconstrained, real-world conditions. It addressed the long-standing issue of annotation inconsistency across datasets by employing a semi-automatic annotation tool, resulting in more accurate and standardized facial keypoints. By consolidating and re-annotation widely used datasets with a common landmark scheme, the 300-W Challenge enabled fair comparison across competing models. This approach of rigorous dataset curation and benchmarking serves as an important precedent for our project, which similarly emphasized the importance of consistent and biologically meaningful keypoint definitions for non-human species. Inspired by 300-W, we adopt standardized keypoint schemes and species-specific evaluation subsets to more accurately assess pose estimation performance in antelopes and beyond.

## III. Background

Pose estimation is the technique of identifying and mapping key body points, like joints, to analyze a subject's movement. In this study, we focus on antelopes as our primary species of interest for pose estimation. Antelopes have a more rigid skeletal structure compared to primates, leading to more predictable movements. This makes them an excellent starting point for developing and testing pose estimation models for animals.

By applying pose estimation to antelopes, we can gain valuable insights into their interactions with predators, particularly how they adjust their posture and behavior in response to threats. This could improve our understanding of how these animals survive in the wild. Moreover, pose estimation has practical applications in managing livestock, such as detecting changes in posture due to illness and enabling more effective monitoring of herd health over time.

We use the AP-10K [1] dataset, a large-scale animal pose estimation benchmark that includes 10,000 labeled images and 50,000 unlabeled images covering 53 species. Among these, the dataset provides 200 labeled images of antelopes.

All of our experiments are conducted using RTMPose [5], a pose estimation model initially developed for detecting multiple people in human-centric datasets. Its capacity to generalize to scenarios involving multiple subjects makes it well suited for animal studies, particularly for species that tend to move in groups, such as antelopes.

## IV. Data Focusing Based on Similiar Species

### A. Overview

When training a keypoint estimation model, the data chosen to train it is a crucial step. [Link to Open Monkey Pose], and [Link to Open Ape Pose] proved the seemingly obvious statement that to optimize a keypoint estimation model for performance on a particular species, training on similar animals to that species generates better keypoint estimation models than training on animals that are not similar to that species. However, previous work like [LINK open monkey], and [LINK open ape] does not answer the obvious question of how does one determine which animals are "similar" to a given species, and what does "similar" mean/in what respect do the "species" need to be similar to produce better performing models when trained from images of those species.

Answering these two essential questions to elucidate insights on how to shrink the amount of data used to train keypoint estimation models while preserving or increasing model performance became the primary goal of these species similarity efforts. Elucidating these insights is important to applying the task of animal keypoint estimation to real-world scenarios and data (in-the-wild camera trap data), as generally, data is much more sparse, and so having a comprehensive way to determine exactly what data is necessary for training a keypoint estimation model is essential.

Last semester, the team focused on determining species similarity to antelopes based on taxonomical similarity, with the two main points of comparison being the Bovidae vs. Bovidae + Cervidae groups of data (normalized for the number of images), given that Antelopes are part of the Bovidae group. The result that the model trained on Bovidae + Cervidae data outperformed the model trained on just Bovidae data gave credence to the point that quantifying the most effective way to train a keypoint estimation model based on the application was not trivial and intuitive.

Thus, this semester the team focused our efforts on determining similar species to Antelopes to train RTMPose models on based on morphology, as keypoint estimation is inherently a visual task, so we hypothesized that data generated based on morphological metrics could produce better models than models from taxonomical-based data. We chose three quantitative metrics to explore: centroid variation, limb ratios, and DINO+SD, as they each covered a different portion of the morphology of animals, and so they give an accurate part of the morphology that is most significant for creating keypoint estimation training data. Finally, we compared the models trained on data from these morphological metrics with models trained from 10 random species as a baseline and the 10 most similar species chosen by humans to determine if the quantitative metrics can outperform human preference.

### B. Methods

*1) Centroid Variation:* In order to determine species similarity for keypoint based tasks, we use a centroid-based keypoint approach, which aims to analyze the range of motion of the animals by computing their keypoint variations using a centroid as a relative measure. In order to compute these centroid varations, we first developed a script that processes keypoint annotations from the AP-10K [1] dataset and extracts variation metrics. This is done so by loading in the keypoints from the annotation files, where each keypoint is represented as:

$$[x, y, \text{visibility}]$$

After this, we only keep the keypoints that are visible (setting non visible keypoints to $-\infty$), and using these keypoints the centroid is calculated as the mean of all visible keypoints. Once the centroid is computed using the keypoints, we are able to compute the Euclidean distance of each keypoint from the centroid. Finally, for each image, we compute the centroid varation metric which is computed as:

$$V_i = \frac{d_i}{\text{mean}(d)}$$

where $d_i$ is the distance of keypoint $i$ from the centroid, and $\text{mean}(d)$ is the mean distance across all visible keypoints. If the mean distance is zero (no movement), zero variation is returned. Variations are then stored per keypoint are saved in species-specific CSV files. These variations are averaged across all images within each species to provide one variation vector per species of CSV.

Once the CSV files are correctly stored, we use another script to read the vector associated with each species. We also replace $-\infty$ values with NaN to ensure all non visible keypoints are removed from the final vector. After these steps, we then construct a feature matrix where each species is represented by its keypoint variation vector. We then use these vectors to compute cosine similarity between two species (specifically species similarity with respect to antelopes):

$$\text{Similarity}(S_a, S_b) = \frac{\mathbf{V}_a \cdot \mathbf{V}_b}{\|\mathbf{V}_a\|\|\mathbf{V}_b\|}$$

where $S_a$ represents the antelope vector, and $S_b$ is any other species. We then extract the top 10 species with the highest similarity to antelopes to conduct our species similarity expirments. Species with high cosine similarity to antelope likely share similar movements based on keypoints, making them suitable reference species for transfer learning in keypoint-based tasks.

*2) Limb Ratios:* For the limb ratio we first need to define the concept of a "skeleton", which we define as the line between certain keypoints such as the line between the left and right eye keypoints, or the line between the nose and neck keypoints. We used the side view images of the different species in AP-10K [1] for statistics and calculations. We tried to normalize each skeletal segment length $L_{seg}$ is by the specimen's bounding box height $H_{bb}$:

$$\hat{L}_{seg} = \frac{L_{seg}}{H_{bb}}$$

For each species $S_i$, we compute mean normalized segment lengths across all specimens:

$$\mathbf{V}_i = \left[\mu(\hat{L}_1),\ \mu(\hat{L}_2),\ \ldots,\ \mu(\hat{L}_{17})\right]$$

We will get a 17-dimensional vector which is because we have 17 skeletons in total. To find the species that are most similar to the antelope, we use the skeleton vectors of each species and the skeleton vectors of the antelope to compute the cosine similarity between the two species

$$\text{Similarity}(S_a, S_b) = \frac{\mathbf{V}_a \cdot \mathbf{V}_b}{\|\mathbf{V}_a\|\|\mathbf{V}_b\|}$$

where $S_a$ represents the antelope and $S_b$ is another species. We calculate the similarity of the antelope to any other

(A) I think you have a concept in your head that you haven't explained yet. It may make sense to add a paragraph early in this subsection to explain, intuitively, what this approach is trying to measure. The notion of "movement" ties in with this. (Note - I think this parenthetical connects to your intuitive "what is centroid variation measuring?"

(see the description at the end of this subsection → this can be expanded AND copied/moved to the beginning of the subsection to help the reader understand why you did what you did.